Ascertainment bias in the estimation of the effective population size from genome-wide SNP data

Ulrike Ober[1], Alexander Malinowski[2], Martin Schlather[3], Henner Simianer[1]

[1] Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-University Göttingen, Germany

[2] Institute for Mathematical Stochastics, Georg-August-University Göttingen, Germany

[3] Institute for Mathematics, University of Mannheim, Germany




Introduction

Effective population size N_e is a central parameter in population and quantitative genetics

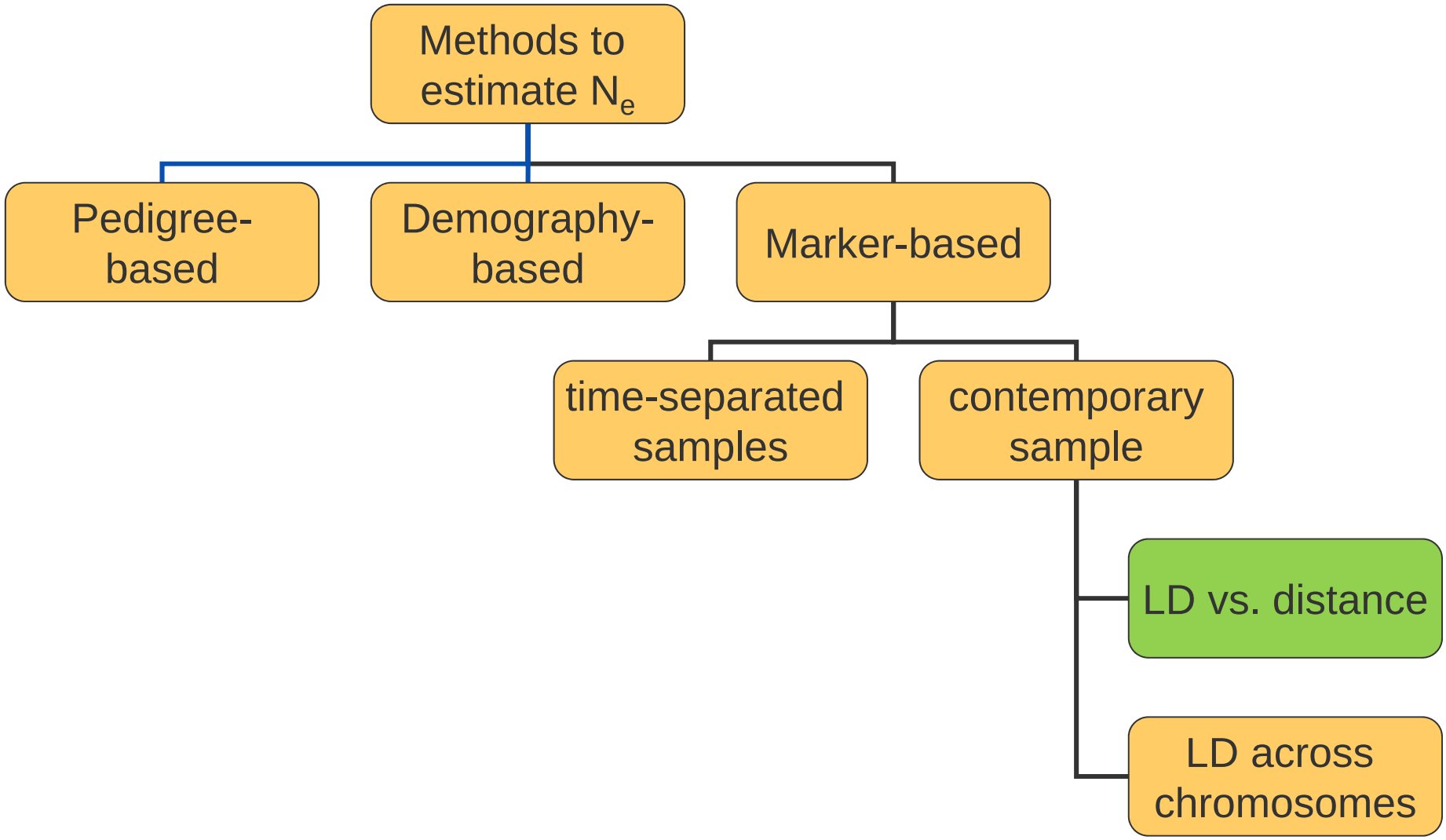
<u>Definition:</u> The effective size N_e of a given real population is the size of a hypothetical ideal population that displays the same characteristics (e.g. inbreeding rate, drift variance, linkage disequilibrium structure) as the real population.

Where does N_e play a role? E.g. for ...

- Development of inbreeding in a closed population
- Definition of conservation priorities
- Accuracy of genomic breeding values $r_{GBV,TBV} = \sqrt{\frac{n_t h^2}{n_t h^2 + \frac{2N_e Lk}{\ln(2N_e L)}}}$

(Goddard et al. 2011)

Introduction



Estimating N_e in a contemporary sample from LD

Sved (1971) $$E(r^2) = \frac{1}{1+4N_e c}$$

where

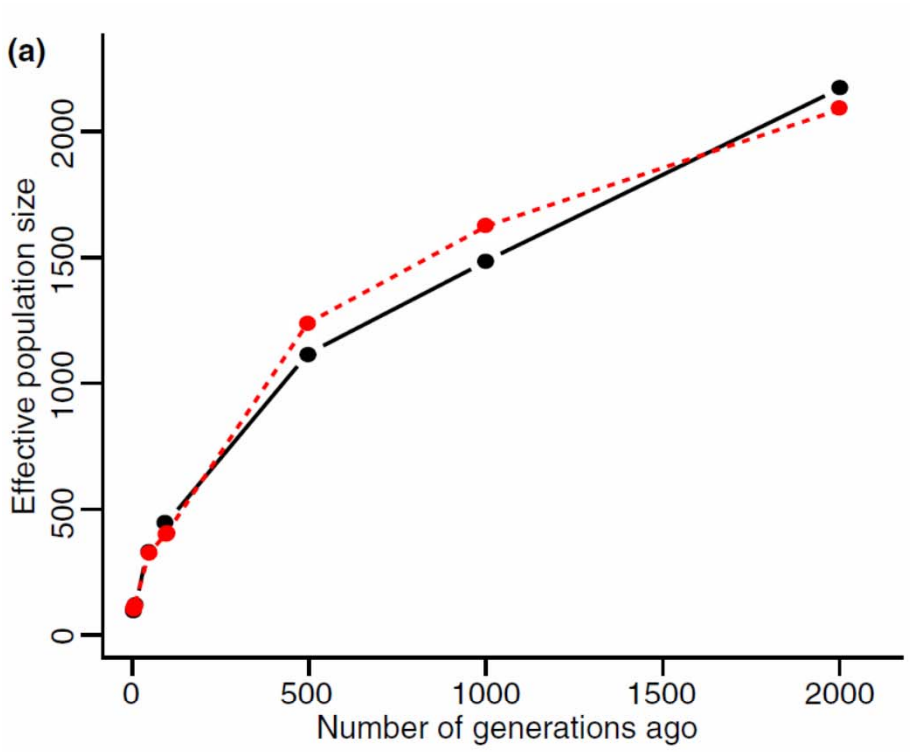
r^2 is the correlation between gametic states at the two loci

c is the distance of loci in Morgan

$$\hat{N}_{e,c} = \frac{1-\left(\overline{r_c^2} - \frac{1}{2n}\right)}{4c\left(\overline{r_c^2} - \frac{1}{2n}\right)}$$

correction for sample size n (according to Bishop et al. 1975)

N_e of Holstein cattle (Qanbari et al., 2009)



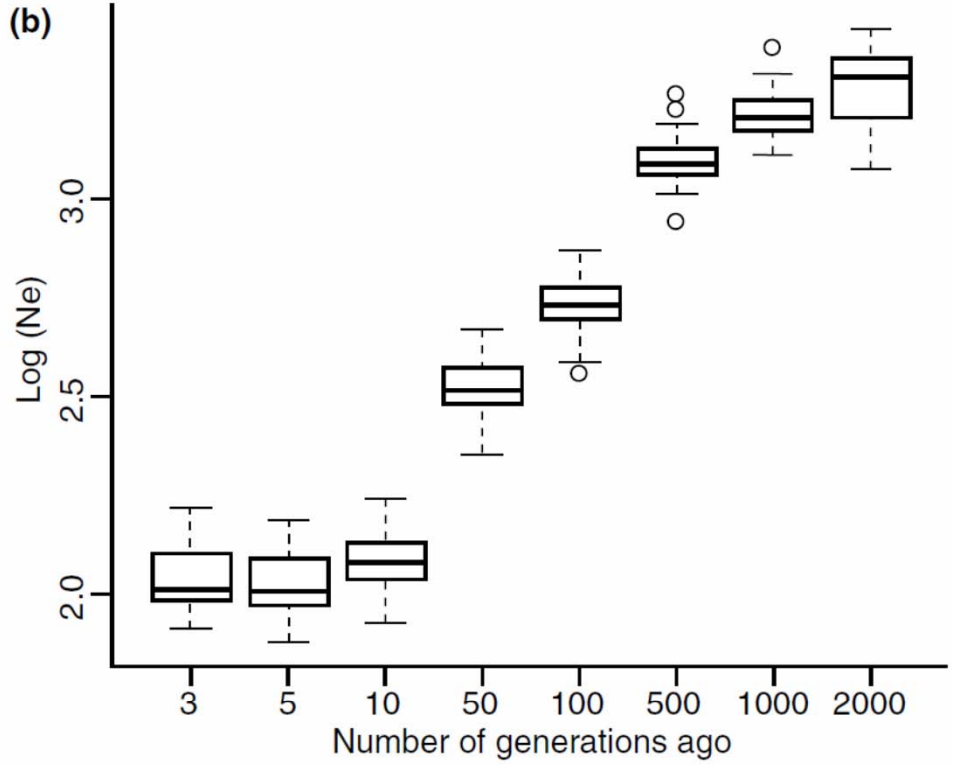


Figure 8 Estimated effective population size over the past generations from linkage disequilibrium data. (a) Dashed and solid lines represent N_e based on estimates of recombination rates and approximate linkage distances respectively. (b) Boxplot representing the trend of $\log_{10}(N_e)$ over time. The variability at each point of time reflects the variation of estimates between the 29 autosomes.

A closer look at Sved‘s (1971) derivation

Sved‘s recursion formula

- development of r^2 from generation t to $t+1$
- between two loci that are c Morgan apart
- in a closed ideal population of size N

$$E(r_{t+1}^2) = \left(1 - \frac{1}{2N}\right)(1-c)^2 E(r_t^2) + \frac{1}{2N}(1-c)^2 \xrightarrow{t \to \infty} E(r_\infty^2) = \frac{1}{1+4Nc}$$

A closer look at Sved‘s (1971) derivation

BAD NEWS...

No mathematically valid derivation for this recursion formula exists.

From John Sved‘s homepage: „This was all introduced in a very messy way, and was not understood by anyone, evidently including myself.“

GOOD NEWS...

Simulation results indicate that the formula works reasonably well

But: the exact recursion depends on allele frequencies

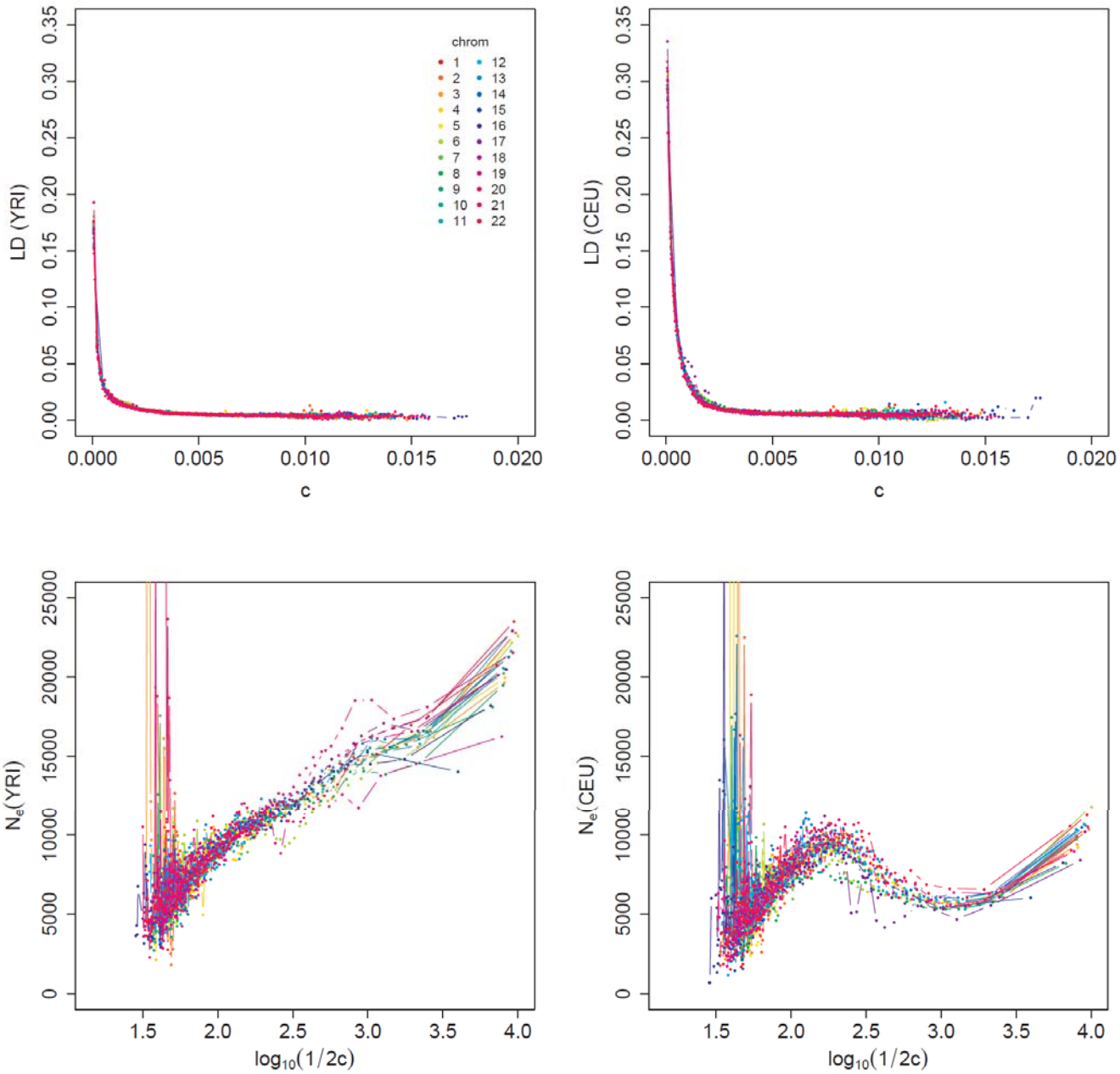
An obvious question: How does the allele frequency spectrum affect the estimates of N_e ?

Data: human Hapmap data (release #27), 22 autosomes

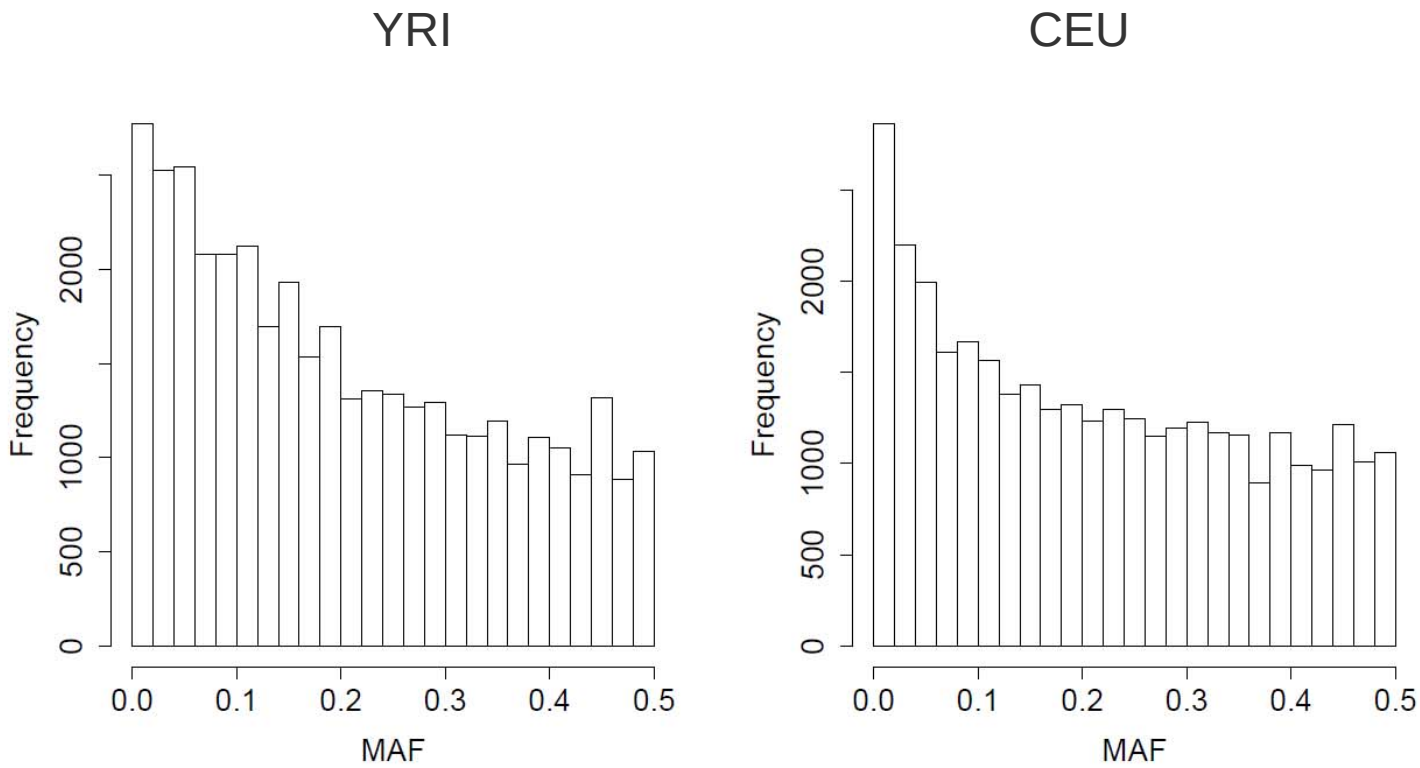
	YRI Yoruba in Ibadan, Nigeria	CEU Western/Northern Europeans from Utah
# of trios	30	30
# of SNPs < 200 kb apart	2.86×10^6	2.56×10^6
# of LD values	702×10^6	563×10^6

YRI

CEU


chrom
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
LD (YRI)
LD (CEU)
c
N_e(YRI)
N_e(CEU)
$\log_{10}(1/2c)$

Minor allele frequency distribution in sequence data



original MAF distribution: 10‘000 SNPs sampled at random

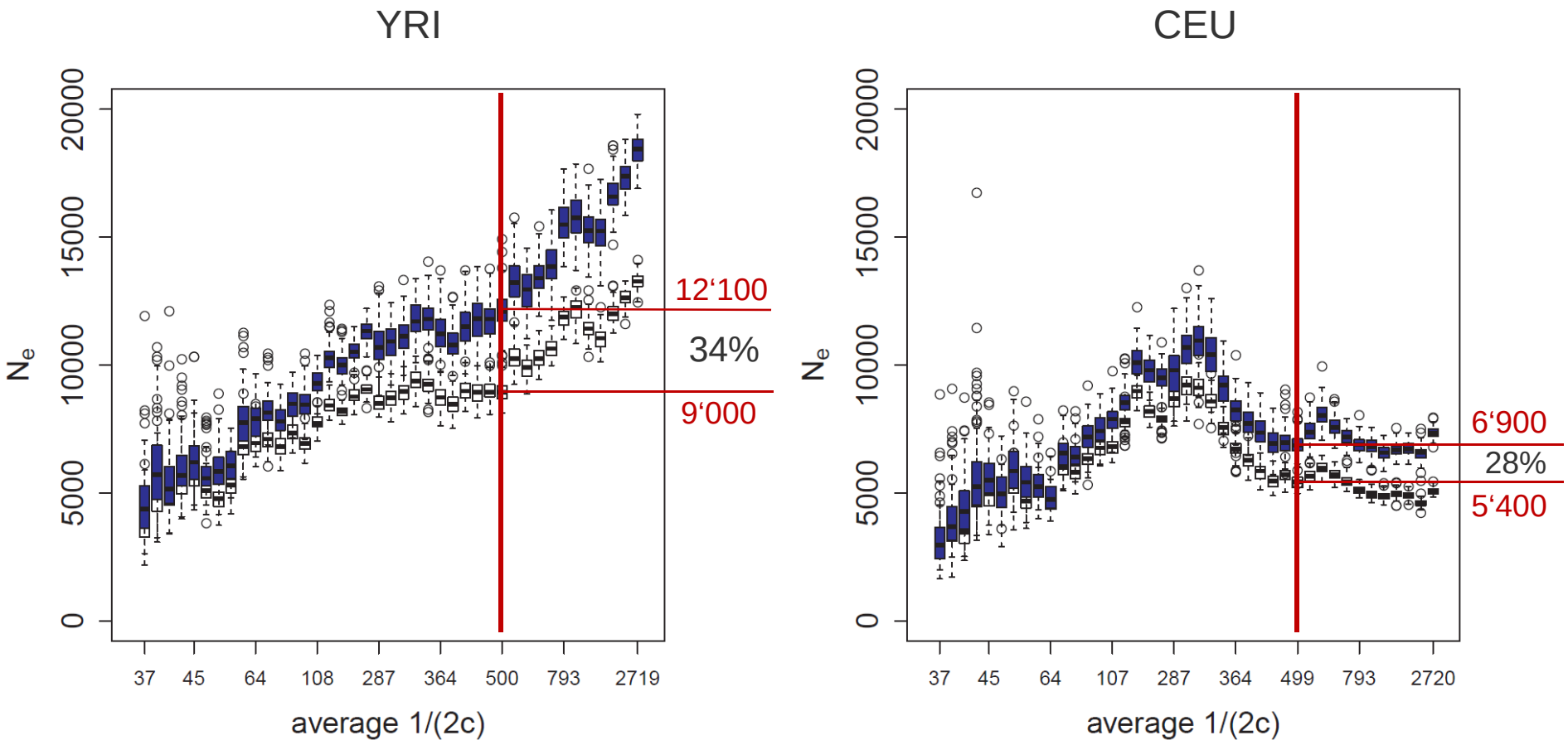
uniform MAF distribution: 1‘000 SNPs sampled at random in each of 10 bins (0.00 – 0.05; 0.05 – 0.10; ... ; 0.45 – 0.50)

In both populations, 100 replicates, results shown for chromosome 22 only

Estimated N_e from different SNP sets

original MAF distribution

imposed uniform MAF distribution



Summary and Conclusions

- Effective population size N_e is a relevant parameter in many areas of population and conservation genetics
- With high density SNP genotypes N_e can be estimated from pairwise LD for different time points in the past
- The underlying recursion formula suggested by Sved (1971) is largely heuristic and lacks a sound mathematical justification, but empirically seems to work reasonably well
- Sved's approach is sensitive to the allele frequency spectrum
- When using a SNP chip with an imposed uniform MAF distribution, historic N_e may be underestimated by ~ 30%
- More methodological research on estimation of N_e from LD is needed

Thank you

This research was funded by

the German Federal Ministry of Education and Research (BMBF)
within the AgroClustEr
"Synbreed - Synergistic plant and animal breeding" (FKZ 0315528C)
in association with the DFG research training group
"Scaling problems in statistics" (RTG 1644)

Bundesministerium
für Bildung
und Forschung

Synbreed

DFG